Cray XD1™ System Overview

Private

S–2429–131

New Features

Cray XD1™ System Overview S–2429–131

This rewrite of *Cray XD1 System Overview* (S–2429) supports the general availability of release 1.3 of the Cray XD1 product. It includes a conversion to the new document format.

Record of Revision

Version	*Description*
1.3.1	October 2005 Supports the general availability of release 1.3 of the Cray XD1 product. Conversion to new document format.
1.3	July 2005 Supports the limited availability of release 1.3 of the Cray XD1 product.
1.2	April 2005 Supports release 1.2 of the Cray XD1 product.
1.1	October 2004 Supports release 1.1 of the Cray XD1 product.
1.0	August 2004 Initial version. Supports release 1.0 of the Cray XD1 product.

Contents

Page

Figures

Preface

The information in this preface is common to Cray documentation provided with this software release.

Accessing Product Documentation

With each software release, Cray provides books and man pages, and in some cases, third-party documentation. These documents are provided in the following ways:

CrayDoc
: The Cray documentation delivery system that allows you to quickly access and search Cray books, man pages, and in some cases, third-party documentation. Access this HTML and PDF documentation via CrayDoc at the following locations:

  - The local network location defined by your system administrator
  - The CrayDoc public website: `docs.cray.com`

Man pages
: Access man pages by entering the `man` command followed by the name of the man page. For more information about man pages, see the `man`(1) man page by entering:

  `% man man`

Third-party documentation
: Access third-party documentation not provided through CrayDoc according to the information provided with the product.

Conventions

These conventions are used throughout Cray documentation:

Convention	Meaning
`command`	This fixed-space font denotes literal items, such as file names, pathnames, man page names, command names, and programming language elements.
variable	Italic typeface indicates an element that you will replace with a specific value. For instance, you may replace *filename* with the name `datafile` in your program. It also denotes a word or concept being defined.
`user input`	This bold, fixed-space font denotes literal items that the user enters in interactive sessions. Output is shown in nonbold, fixed-space font.
[ ]	Brackets enclose optional portions of a syntax representation for a command, library routine, system call, and so on.
`...`	Ellipses indicate that a preceding element can be repeated.
`name`(N)	Denotes man pages that provide system and programming reference information. Each man page is referred to by its name followed by a section number in parentheses. Enter: `%` **`man man`** to see the meaning of each section number for your particular system.

Reader Comments

Contact us with any comments that will help us to improve the accuracy and usability of this document. Be sure to include the title and number of the document with your comments. We value your comments and will respond to them promptly. Contact us in any of the following ways:

E-mail:
docs@cray.com

Telephone (inside U.S., Canada):
1–800–950–2729 (Cray Customer Support Center)

Telephone (outside U.S., Canada):
+1–715–726–4993 (Cray Customer Support Center)

Mail:
Software Publications
Cray Inc.
1340 Mendota Heights Road
Mendota Heights, MN 55120–1128
USA

Cray XD1 Support

Obtain support for the Cray XD1 product in either of the following ways:

Telephone:
1–888–279–2729 (Cray XD1 Customer Support Center)

Through the CRInform website:
http://crinform.cray.com/xd/

> **Note:** Use the contact information provided here if you have a support agreement with Cray. If, however, you have a support agreement with a third-party organization that is a Cray channel partner, contact that organization instead: do not contact Cray directly.

Introduction [1]

This chapter describes the intended audience for this manual and its scope, and lists the related publications.

1.1 Who Should Read This Manual

This manual is intended to introduce the Cray XD1 supercomputer to a technician who is responsible for installing and maintaining it, an administrator who is responsible for commissioning, managing, and monitoring it, and an end user who will run jobs on it.

1.2 Scope of This Manual

This manual describes the main hardware and software components as well as the user environment of the Cray XD1 supercomputer. For more detail about the Active Manager software that is used to manage, monitor, and submit jobs in a Cray XD1 system, see *Cray XD1 System Administration* (2430).

1.3 Related Publications

Refer to the publications listed in Table 1, page 1 for more information on how to use the Cray XD1 computer and the Active Manager software.

Table 1. Related publications

Publication title	**Brief description**
Cray XD1 Release Description (S-2453)	Identifies the main new features and enhancements in a particular release of the product. Includes information about the hardware, embedded software, and Linux based software of the system.
Cray XD1 Site Planning (HR6-6401)	Guidelines on how to plan and prepare a facility for a Cray XD1 installation.

Publication title	**Brief description**
Cray XD1 Hardware Installation and Upgrade (HR6-6402)	Hardware installation and hardware upgrade procedures.
Cray XD1 RapidArray Interconnect Topologies (HR6-6425)	Guidelines on cabling Cray XD1 hardware components in RapidArray interconnect topologies.
Cray XD1 System Administration (S-2430)	System administration and monitoring procedures. Also job submission and management procedures.
Cray XD1 Programming (S-2433)	Development tools on a Cray XD1 system and the application programming interface for the field-programmable gate array (FPGA).

What is the Cray XD1 Supercomputer? [2]

The high-bandwidth, low-latency Cray XD1 system is designed specifically for high performance computing (HPC). This modular, scalable system unifies tens or hundreds of processors into a single, balanced supercomputer.

2.1 Physical Description

This section introduces a Cray XD1 chassis, a Cray XD1 system, and the interprocessor communications in both a chassis and a system.

2.1.1 Cray XD1 Chassis

The base unit of a Cray XD1 system—a Cray XD1 chassis—contains a maximum of 31 commodity and specialized processors:

- Twelve 64-bit AMD Opteron processors—Configured as six two-way or four-way symmetric multiprocessors (SMPs) that run the Cray high-performance Linux operating system.
- Six or twelve RapidArray processors—Hardware, designed by Cray, that processes most communications in a chassis. These processors provide a high-bandwidth, low-latency interface to the high-speed RapidArray interconnect.
- Zero or six FPGA application acceleration processors—Field-programmable gate arrays (FPGAs) that act as coprocessors to the Opteron processors.
- One management processor—Runs software that monitors and controls the health of the chassis. In a multichassis system, management processors communicate over an independent supervisory network.

Note: Each instance of the Linux operating system is a processing node in a Cray XD1 system. The term *node* in this and other Cray XD1 manuals refers to an instance of the operating system and the hardware components that it controls. In the current release, the hardware components in a node include the two-way or four-way Opteron SMP and its associated memory, one or two RapidArray processors, and an optional application acceleration processor.

Each Cray XD1 chassis occupies three vertical units in a standard four-post, 23-in. cabinet. Four fans cool the assemblies and subassemblies within the enclosure, which controls the flow of air through the chassis. Hardware components connect to a main board, which occupies most of the inside bottom area of the enclosure. For more details, see Chapter 3, page 9.

Figure 1, page 4 shows the front view of a Cray XD1 chassis; Figure 2, page 4 shows the rear view.

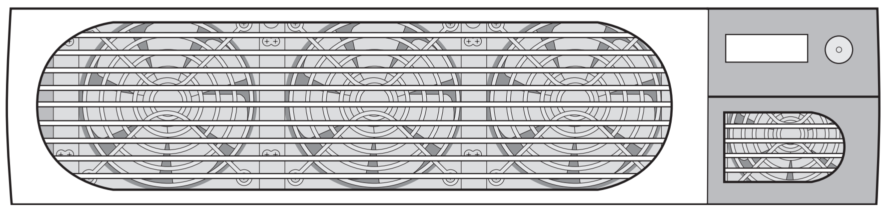

Figure 1. Cray XD1 chassis, front view

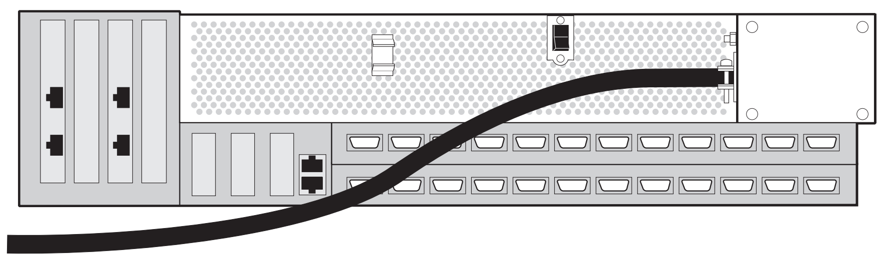

Figure 2. Cray XD1 chassis, rear view

All cooling air is pulled in through the front of the chassis and exhausted at the back. Hinged, perforated fascias on the front of the chassis allow air to enter the enclosure and provide service access to the fans and disk blades. Power and data cables connect to the back of the chassis.

The Cray XD1 chassis can be serviced while it is installed in a cabinet. Fans and disk blades can be replaced while the Cray XD1 is fully cabled and powered on. Other field-replaceable components, including the power supply and compute blades, can be replaced after the chassis is powered off and its cover removed. For maintenance procedures, see the Field Replacement Procedures that are packaged with replacement parts.

2.1.2 Cray XD1 System

Multiple Cray XD1 chassis can be interconnected in various topologies. The term *system* in this document indicates either a stand-alone chassis or a network of interconnected chassis.

In a typical installation, a maximum of 13 Cray XD1 chassis are mounted in a 7-foot cabinet. One or more cabinets are installed in a computer room with a controlled environment. Multicabinet systems may include switch cabinets, which house external switches and other ancillary equipment (such as storage devices). The switch cabinets are positioned between Cray XD1 cabinets to keep cable lengths to a minimum. Raised-floor server rooms, with alternating hot and cold aisles, help cool multicabinet systems; see *Cray XD1 Site Planning* (HR6-6401) for more information.

2.1.3 Interprocessor Communication

Processors communicate within a chassis and between chassis primarily over the high-bandwidth, low-latency RapidArray interconnect. Section 2.2.1, page 5 discusses this interconnect in more detail.

The management processors communicate with the nodes in a chassis over an independent supervisory network. In multichassis systems, the supervisory network (which is interconnected by external Ethernet switches) enables communication among the management processors in all chassis.

2.2 Key Features

This section describes the key features of the Cray XD1 computer.

2.2.1 RapidArray Interconnect

The way in which an HPC system's processors interconnect greatly affects the performance of parallel applications. The Cray XD1 system is based on the Direct Connected Processor (DCP) architecture, which eliminates the major performance and scalability problems of other platforms. Processors and memory connect directly to the high-speed, low-latency RapidArray interconnect, and so avoid PCI bus bottlenecks and shared resource contention. The RapidArray interconnect is the principal mechanism for communications within a Cray XD1 system; it provides reliable transport for all data types.

Each Cray XD1 chassis has either one or two RapidArray fabrics: the main RapidArray fabric and an optional expansion RapidArray fabric. Multichassis systems may include external RapidArray switches. The number of external RapidArray switches depends on the type and size of the physical topology. Direct-connect topologies, in which every chassis has a direct link to every other chassis in the system, have no external switches.

For more details about the RapidArray interconnect and for instructions on how to manage it, see *Cray XD1 System Administration* (S-2430). For instructions on how to cable various standard topologies, see *Cray XD1 RapidArray Interconnect Topologies* (HR6-6425).

2.2.2 Application Acceleration

Each Cray XD1 chassis can include six optional FPGA application acceleration processors, which can be programmed to accelerate computationally intensive and highly repetitive segments of code. Algorithms, or portions of algorithms, can be compiled into hardware, which enables the application accelerator processor to act as a coprocessor to the Opteron processor. The FPGA application acceleration processor is particularly well suited to data stream processing, bit manipulation, integer operations, fabric computing, and random number generation.

The FPGA application acceleration processors are based on the Xilinx Virtex II Pro FPGA. They are optional components on the expansion module. For more information, see Section 3.6, page 22.

2.2.3 Reliability and Availability

The Cray XD1 computer proactively monitors, configures, and manages system hardware and software components. For instance, the system continuously monitors hundreds of critical hardware functions. It assigns IP addresses automatically when the system is commissioned or expanded. It automatically reboots nodes that fail a sanity check. If a node fails, a replacement node can take over without administrator intervention. In summary, many aspects of system administration are automated, which reduces system downtime as well as the total time and effort required to manage the system.

The embedded management processor runs dedicated software—the Active Manager hardware supervisory subsystem—on a real-time operating system and uses an independent supervisory network to monitor, manage, and control the system. The ability of the hardware supervisory subsystem to reduce the incidence of failure and to work around failures when they do occur helps to make the Cray XD1 a highly reliable and available system.

2.2.4 Flexibility

The Cray XD1 computer supports a wide variety of code and easily adapts to increasing workloads.

2.2.4.1 Standards-Based

The Cray XD1 is based on the following standards; therefore, it supports a wide range of commercial and open-source HPC applications:

- x86-64 instruction set—Part of the Opteron processor.
- Linux—Broadly accepted in the HPC user community for its open source and multivendor interoperability. See also Section 4.1, page 29.
- Message Passing Interface (MPI)—A popular parallel programming model. See also Section 4.4, page 35.

2.2.4.2 Support for 32- and 64-bit Computing

Because the Cray XD1 computer uses the AMD 64-bit Opteron processor (single- or dual-core), it is compatible with both 32- and 64-bit applications. This means that end users do not need to rewrite or recompile existing x86 code to move to a Cray XD1 system.

2.2.4.3 Scalability

The scalability of parallel programs is limited by the bandwidth and latency of the interconnect. The high-bandwidth, low-latency RapidArray interconnect offers better performance and greater scalability than alternative interconnects, such as Gigabit Ethernet. In addition, the Cray XD1 system management features enable administrators to easily grow their systems without significant system downtime.

2.2.5 Single-system Control

Significant numbers of system failures result from administrative challenges—system complexity, software and hardware incompatibilities, and the manual nature of administrative tasks such as installation, configuration, upgrades, and troubleshooting. The task-centric Active Manager software alleviates these problems: it provides administrators with a single point of control for the whole system through either a graphical user interface (GUI) or command-line interface (CLI). See Section 4.3, page 31 for more details.

Cray XD1 Hardware Components [3]

This chapter describes the major Cray XD1 hardware components and how they interoperate.

3.1 Summary of Hardware Components

This section provides a brief description of the main Cray XD1 hardware components; see Figure 3, page 9 for an exploded view of a chassis.

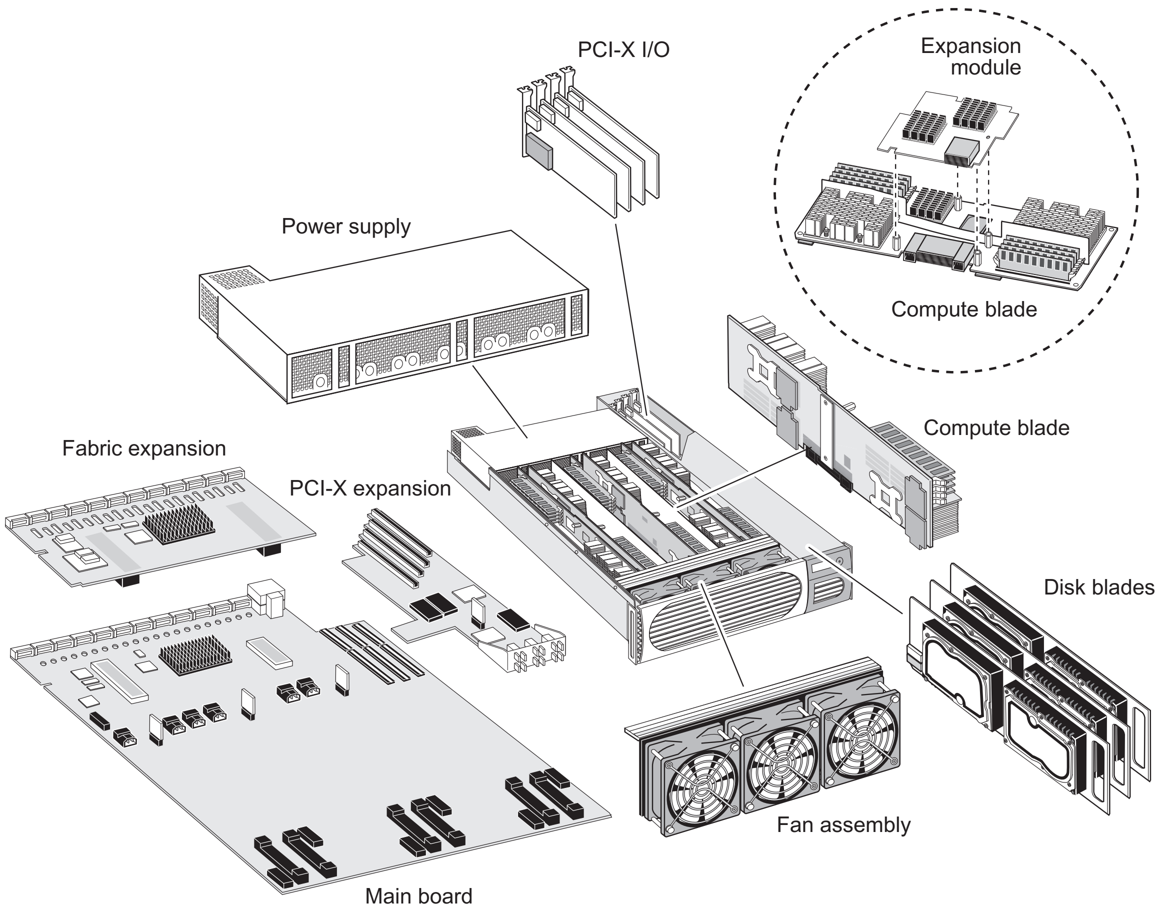

Figure 3. Exploded view of a Cray XD1 chassis

Table 2, page 10 summarizes the main hardware components of a Cray XD1 chassis. The remainder of this chapter provides more detail about each component. The base configuration is an enclosure with left and right-side fascias, a main board, six compute blades, four fans, and a power supply. Other components are optional.

Table 2. Summary of hardware components

Component	**Number per chassis**	**Description**
Enclosure	1	The enclosure protects the assemblies and subassemblies of a Cray XD1 chassis and regulates airflow. The physical interfaces and status LEDs are visible on the outside of the enclosure. An LCD panel in on the right-side fascia.
Main board	1	The main board includes the management processor, the main RapidArray switch, 24 internal RapidArray links, an internal Ethernet switch, and connectors for other components.
Compute blades	6	Each compute blade includes a dual-Opteron SMP (single- or dual-core), DIMMs (DDR SDRAM), and a RapidArray processor. The latter provides two 2–GBps links to the main RapidArray switch.
Expansion modules	0 or 6	Optional. The expansion modules attach to each compute blade. They provide an additional RapidArray processor and two additional internal RapidArray links per compute blade, and an optional FPGA application acceleration processor. If the expansion modules are present, a chassis has a total of 12 RapidArray processors and may also have 6 FPGA application acceleration processors. To make use of the additional RapidArray components, a chassis must also include the fabric expansion card, described below.
Fabric expansion card	0 or 1	Optional. The fabric expansion card provides a chassis with a second RapidArray switch (known as the expansion switch). Without this card, a chassis has 12 external RapidArray ports; with this card, 24. The fabric expansion card is required only if the six compute blades have the optional expansion modules, described above.

Component	Number per chassis	Description
PCI-X expansion card	0 or 1	Optional. The PCI-X expansion card provides four PCI-X buses (one slot each) for NICs and Fibre Channel cards. It also includes connectors for up to three disk blades. Chassis that do not have a PCI-X expansion card may have a disk expansion card (described below) instead.
Disk expansion card	0 or 1	Optional. The disk expansion card may be present in chassis that do not contain a PCI-X expansion card. It has connectors for up to three disk blades.
PCI-X cards	0 to 4	Optional. Each chassis may have one to four PCI-X cards. NICs (for LAN access) and Fibre Channel cards (for storage access) are supported. Both types of PCI-X cards share the same set of four PCI–X slots.
JTAG interface card	0 to 3	Optional. The JTAG interface card tests the FPGA application acceleration processor’s integrated circuits. It connects to a high-speed I/O slot on the main board.
Disk blades (serial ATA)	1 to 3	Disk blades house either one or two serial ATA disks, so each chassis has one to six disks. Each disk provides directly attached storage to a particular node. The disk blades connect to the PCI-X expansion card. They can be replaced while the chassis is powered on.
Power supply	1	Each chassis has a single power supply that is designed with six separate channels (one per node) to tolerate failures.
Fans	4	Each chassis has four fans. All are accessible behind the fascias and can be replaced while the Cray XD1 system is powered on.
Blanking panels		Various blanking panels maintain the EMI seal of the enclosure when optional components are not present. They also help to regulate cooling.

3.2 Chassis Enclosure

This section describes the Cray XD1 chassis enclosure.

3.2.1 Dimensions

The Cray XD1 chassis occupies three vertical units in a standard 23-in. cabinet; a maximum of 13 chassis fit in a seven-foot cabinet. See Table 3, page 12 for the dimensions of a chassis.

Table 3. Cray XD1 chassis dimensions

Characteristic	Dimension (in.)
Height	5.25 (3 VU)
Width	21.50
Depth	36.00

3.2.2 Power Switch

The power switch, located on the rear of the chassis, switches power between On and Standby modes. To remove power completely from the system, disconnect the power cord plug from the power source. For information on the power supply, see Section 3.13, page 26.

3.2.3 Ports

The ports on a Cray XD1 chassis vary depending on the installed options; see Table 4, page 12.

Table 4. Externally visible ports

Port type	Qty	Purpose	Comment
100 Mbps Ethernet	2	• Interchassis management communications • Service access	System management is done over an independent supervisory network. These ports provide physical connections to the supervisory network. The supervisory network must be separate from the site LAN.

Port type	Qty	Purpose	Comment
RapidArray	12 or 24	Interchassis communications	In the base configuration, 12 external RapidArray ports are available. 24 are available if the chassis is configured with a fabric expansion card.
NIC	Depends on NIC type	LAN and NAS access	Up to four NICs may be installed in the PCI-X slots.
Fibre Channel	0 to 8	External storage access	Up to four single-port or dual-port Fibre Channel cards may be installed in the PCI-X slots.
JTAG	0, 2, 4, 6	JTAG interface	Used for debugging FPGA application acceleration processors.

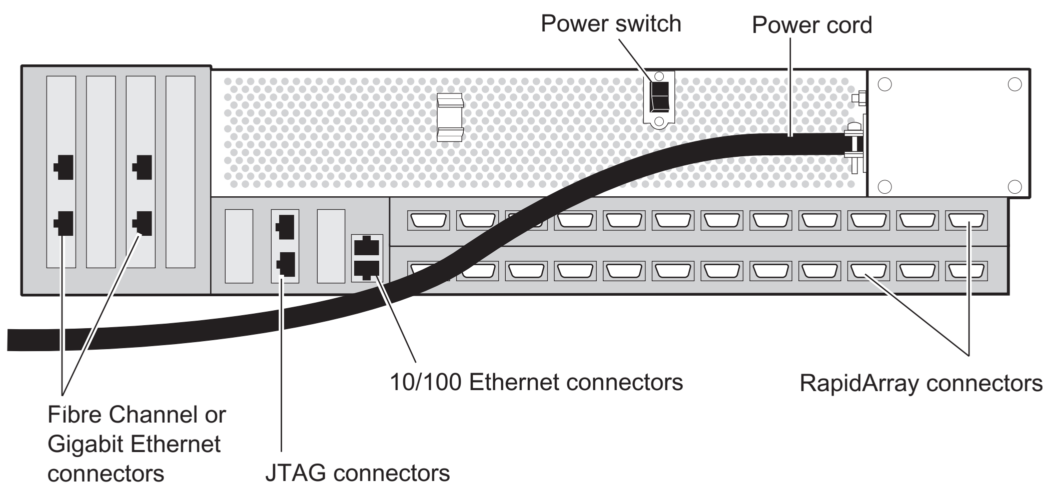

Figure 4. External ports on the rear of the Cray XD1 chassis

3.2.4 LCD Panel

External storage access

The LCD panel on the right-side fascia has a two-row display. Each row displays a maximum of 16 characters at a time:

- The top row displays the IP address of the management processor on the supervisory network.
- The bottom row can display a maximum of 80 characters: it usually displays the ID or name of the chassis. During hardware maintenance activities, this row displays information to help identify the component to be replaced. Administrators can also post messages here. The display scrolls horizontally if the status information exceeds 16 characters.

3.2.5 Front-Panel Button

The front-panel button, next to the LCD on the right-side fascia, is for use by Cray service personnel only.

3.2.6 Chassis Identification Sticker

A chassis identification sticker behind the right-side fascia provides the following information:

- Chassis ID in hexadecimal notation.
- Default IP address of the management processor (on the supervisory network). This IP address may change when the system is commissioned; if the IP address on this sticker and the IP address on the LCD panel are different, the LCD shows the correct address.
- Chassis serial number in decimal notation. This serial number is equivalent to the chassis ID that appears in the Active Manager user interfaces.

3.2.7 Status LEDs

This section describes the various LEDs on the exterior of the chassis.

3.2.7.1 Power Supply LEDs

The power supply LEDs are on the rear of the chassis; see Figure 5, page 15.

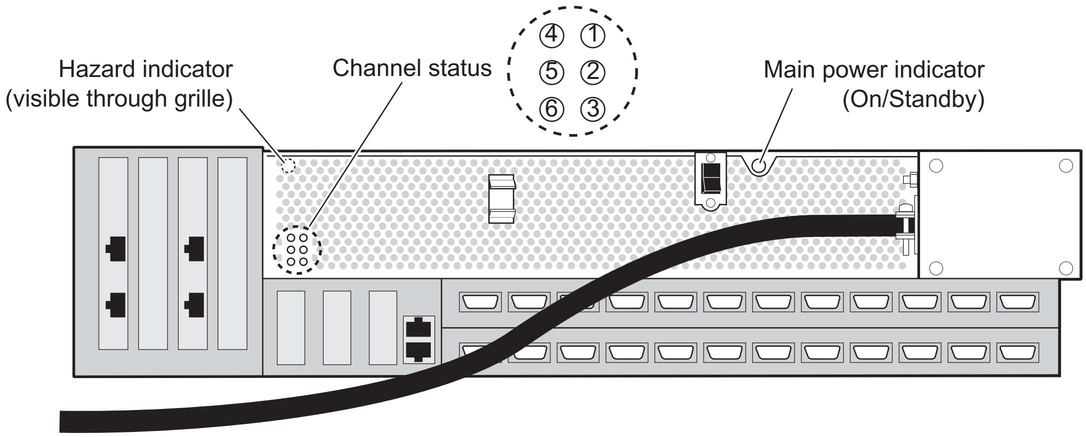

Figure 5. Power supply LEDs

Table 5, page 15 describes the power supply LEDs.

Table 5. Description of the power supply LEDs

LED	Location	Qty	Color	Description
Main power indicator Shows the On / Standby status	Rear; to the right of the power switch	1	None	No power to the chassis; the power cord is disconnected from the power source
			Amber	Power cord is connected to the power source; power switch is turned to the Standby position
			Green	Power cord is connected to the power source; power switch is turned to the On position

LED	Location	Qty	Color	Description
Channel status Each power channel provides power to a node (The numbering on Figure 5, page 15 shows the mapping of LED to node)	Rear; lower left of the power supply	6	None	No power to the channel
			Amber	Channel is in standby and provides low power to support a subset of the internal circuitry Appears briefly during the chassis power-on sequence: during the normal operation of the system, an amber LED usually indicates a component failure on the node
			Green	Active power output for the channel
			Red	Power channel is in a fault state A red LED usually indicates a fault in the power supply itself
Hazard indicator Visible through the grille	Rear; top left of the power supply	1	None	Hazardous voltages are not present
			Red	Hazardous voltages are present **Note:** This LED is on whenever the chassis is connected to a power source. After the chassis is disconnected from the power source, the LED remains on until the hazardous voltage discharges.

3.2.7.2 Link LEDs

The link LEDs are visible on the rear of the chassis, adjacent to each port. Table 6, page 17 describes the link LEDs. Each RapidArray and Ethernet link has two LEDs that indicate its logical and physical statuses.

Table 6. Description of link LEDs

LED	Location	Qty	Color	Description
RapidArray logical link status	Rear; adjacent to each RapidArray port	12 or 24	None	Link inactive
			Amber; steady	Link active; no traffic
			Amber; flickering	Link active; packets being transmitted or received
RapidArray physical link status	Rear; adjacent to each RapidArray port	12 or 24	None	Physical link down
			Green	Physical link up
Ethernet physical link status	Rear; adjacent to each Ethernet port	2	None	Physical link down
			Green; steady	Physical link up; no activity
			Green; flickering	Physical link up; activity
Ethernet logical link status	Rear; adjacent to each Ethernet port	2	These LEDs are not used in the current release	

3.2.7.3 Disk Blade LEDs

Each disk blade has two LEDs: the top is for Disk A; the bottom, for Disk B. See Table 7, page 17 for a description of the LEDs; see Figure 14, page 26 for a diagram that shows the disk blade LEDs. The disk blades are accessible behind the small fan on the front of the chassis.

Table 7. Disk blade LEDs

Color	Description
None	No power to disk
Amber	Disk uninitialized
Red	Disk out of service
Green	Disk in service

3.3 Main Board

The main board occupies most of the bottom area of the enclosure; see Figure 6, page 19. Its principal features are as follows:

- Management processor, which runs management software that uses an independent supervisory network to monitor and actively manage the state of a chassis.

  The management processor is a 400 MHz AMD AU1000 processor that runs the MQX real–time operating system.

- Internal Ethernet switch, for use in the supervisory network.

- Connectors for each of the six compute blades. As Figure 6, page 19 shows, these connectors are numbered 1 through 6 (left to right, as viewed from the front of the chassis).

- Connectors for each of the six optional expansion modules.

- Internal 24-port RapidArray switch (the main switch) and internal RapidArray links.

- Connectors for the optional fabric expansion card, which provides a second RapidArray switch (the expansion switch) and additional RapidArray links.

- Three high-speed I/O slots. As Figure 6, page 19 shows, these slots are numbered 1 through 3 (left to right, as viewed from the rear of the chassis). Each has two HyperTransport links to a pair of nodes: slot 1 connects to nodes 5 and 6, slot 2 connects to nodes 3 and 4, and slot 3 connects to nodes 1 and 2.

  Note: This I/O slot–to–compute blade mapping does not apply to JTAG interface cards.

  Slot 1 is the interface for the PCI-X expansion card. Other slots are available for use by JTAG interface cards.

- Connectors for the power supply. The leftmost connector is the receptacle for the power supply control wire. The remaining six are for each of the power channels.

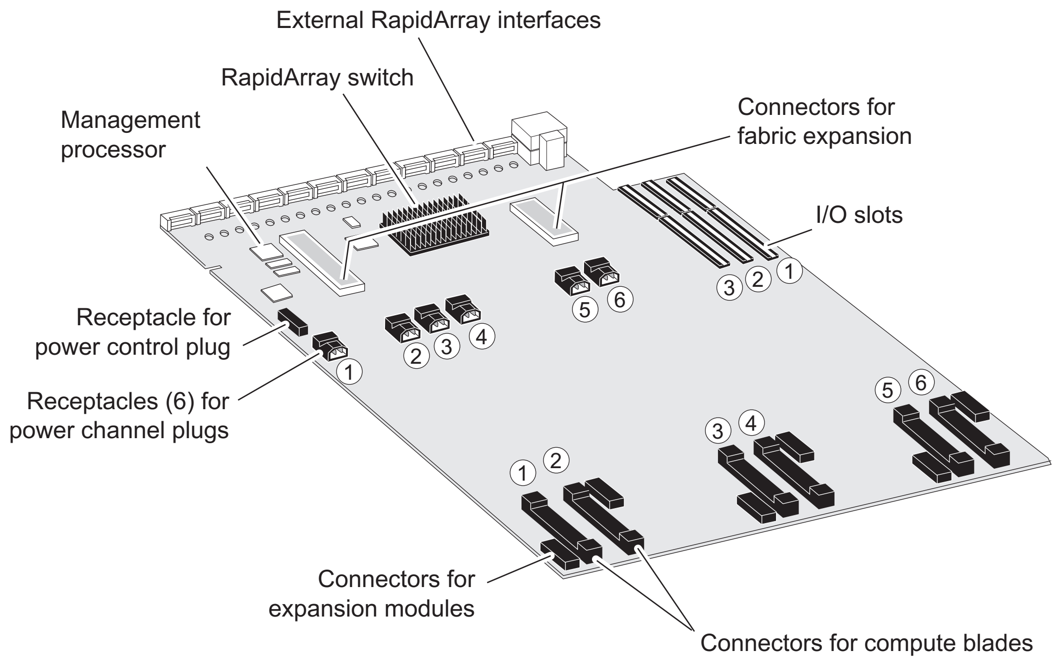

Figure 6. Main board

3.4 Compute Blades

Each Cray XD1 chassis has six compute blades. The compute blades are numbered 1 through 6, left to right, as viewed from the front of the chassis (this numbering also applies to nodes). Each compute blade contains two Opteron processors that are configured as a two-way SMP, four to eight DIMMs, and a RapidArray processor; see Figure 7, page 20.

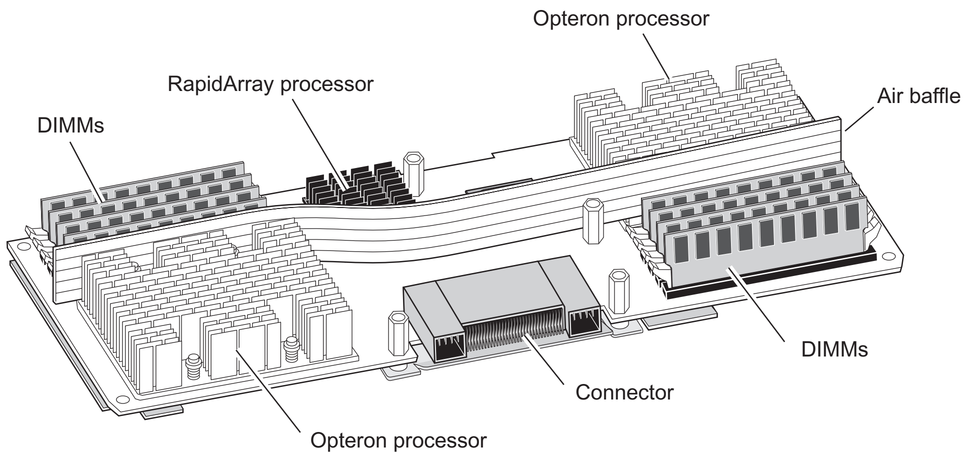

Figure 7. Compute blade

Compute blades may also include an optional expansion module (see Section 3.5, page 21), which provides a second RapidArray processor and an optional field-programmable gate array (FPGA) application acceleration processor.

The compute blades connect to the internal RapidArray switch at four GBps: two GBps each for transmit and receive (both transmit and receive can operate simultaneously). With the optional expansion modules and fabric expansion card, each compute blade connects to the RapidArray switches at eight GBps: four GBps each for transmit and receive.

3.4.1 Opteron Processors

Each compute blade has two 64-bit Opteron processors configured as a SMP. Various processor clock speeds are available; single-core and dual-core models are available. Each Opteron processor has its own caches, its own memory, and its own path out to memory.

3.4.2 Memory

Each compute blade has a maximum of eight dual in-line memory modules (DIMMs)—a maximum of four per Opteron processor—that are available in various sizes and speeds and provide a maximum of 16 GB of DDR SDRAM per node.

3.4.3 RapidArray Processors

Each compute blade has one RapidArray processor and the option of a second one with the addition of the optional expansion module. These processors, designed by Cray, provide the interface between the Opteron processors and the internal RapidArray links. Each RapidArray processor has two bidirectional 2–GBps links to the internal RapidArray switch.

The RapidArray processor has the following features:

- Enables messaging that goes directly from user memory to user memory; bypasses the Linux kernel to access the Opteron processor's memory
- Optimizes short message transmission for low latency and long message transmission for high bandwidth
- Maintains clock synchronization across the entire Cray XD1 system
- Provides reliable transport

3.5 Expansion Modules

The optional expansion modules provide each node with a second RapidArray processor, two additional Rapid Array links, and an optional FPGA application acceleration processor (described in the following section). The expansion modules connect to each compute blade and also directly to the main board. The view in Figure 8, page 21 shows the connector for the main board; the connector for the compute blade is on the underside.

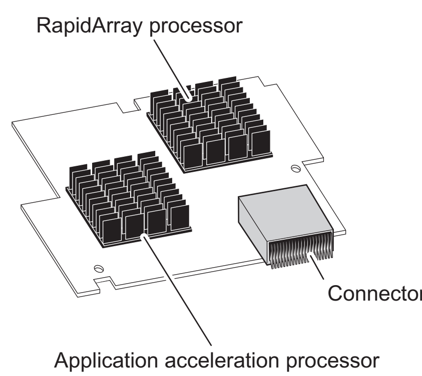

Figure 8. Expansion module

3.6 FPGA Application Acceleration Processor

An expansion module may include an FPGA application acceleration processor, which is available in various sizes and speed grades. This processor is a Xilinx FPGA that users can program to accelerate computationally intensive and repetitive algorithms. The FPGA application acceleration processor has a 3.2-GBps connection to the RapidArray processor.

The FPGA application acceleration processor has 4 banks of SRAM, which function as a high-performance cache, with an aggregate transfer rate of 12.8 GBps. The RapidArray processor can transfer data between Opteron memory and the FPGA application acceleration processor SRAM without interrupting application software or the Linux kernel. For more information on the application acceleration processor, see *Cray XD1 FPGA Development* (S-6400).

3.7 Fabric Expansion Card

The optional fabric expansion card, shown in Figure 9, page 22, provides a Cray XD1 chassis with a second 24–port RapidArray switch (the expansion switch) that doubles its internal switching capacity. It provides 12 additional external RapidArray ports at the back of the chassis, for a total of 24.

The fabric expansion card connects to the main board.

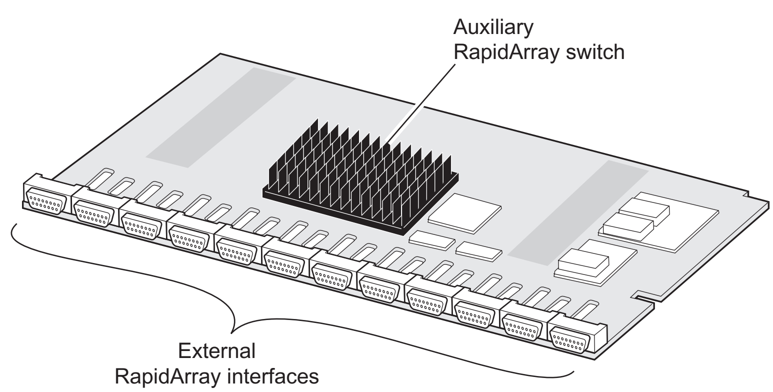

Figure 9. Fabric expansion card

3.8 PCI-X Expansion Card

The optional PCI-X expansion card provides four PCI-X buses for network interface cards (NICs) or Fibre Channel cards and three connectors for disk blades; see Figure 10, page 23. The PCI–X expansion card connects to the first of three high-speed I/O slots on the main board. This slot has HyperTransport connections to compute blades 5 and 6.

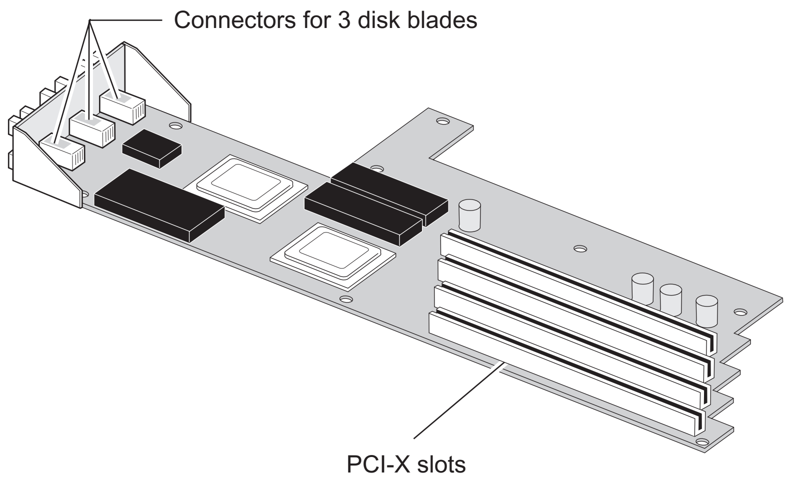

Figure 10. PCI-X expansion card

3.9 Disk Expansion Card

Chassis that do not include a PCI-X expansion card (described above) may have a disk expansion card instead. It includes connectors for up to three disk blades; see Figure 11, page 23.

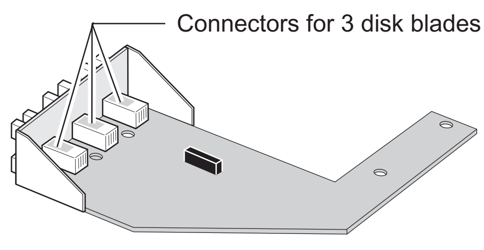

Figure 11. Disk expansion card

3.10 PCI-X Cards

Cray supports NICs and Fibre Channel cards for the Cray XD1 system. Both card types connect to slots on the PCI-X expansion card.

3.10.1 Network Interface Cards

A Cray XD1 chassis can have a maximum of four NICs to provide LAN or NAS access. These cards connect to slots on the PCI-X expansion card; see Section 3.8, page 23. NICs share the four available PCI-X slots with any required Fibre Channel cards.

See *Cray XD1 System Administration* (S-2430) for instructions on providing a system with LAN access.

3.10.2 Fibre Channel Cards

A Cray XD1 chassis has a maximum of four Fibre Channel cards , typically to provide access to storage devices. Fibre Channel cards connect to PCI–X slots on the PCI-X expansion card; see Section 3.8, page 23. Fibre Channel cards share the four available PCI-X slots with any required NICs.

3.11 JTAG Interface Cards

The optional JTAG interface card tests the integrated circuits on the FPGA application acceleration processor; see Figure 12, page 24.

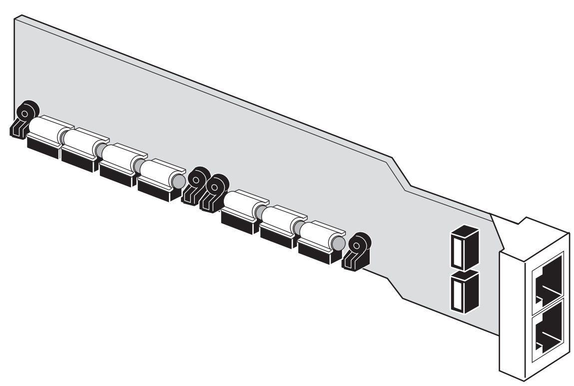

Figure 12. JTAG interface card

These cards connect to the high-speed I/O slots on the main board. The ports on a JTAG interface card are numbered 1 and 2, from top to bottom; see Figure 13, page 25.

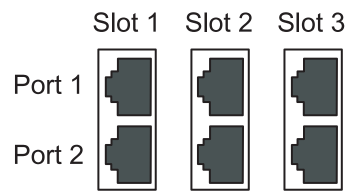

Figure 13. Ports on the JTAG interface card

> **Note:** This figure shows three JTAG interface cards; there may be fewer in a chassis. A Cray XD1 chassis can have a maximum of two JTAG interface cards if the PCI–X expansion card is present because the PCI–X expansion card occupies I/O slot 1.

For more information on the JTAG interface card and for information on programming the FPGA application acceleration processor, see *Cray XD1 FPGA Development* (S-6400).

3.12 Disk Blades

Each Cray XD1 chassis has one to three disk blades; each disk blade holds one or two high-speed, 3.5-in. serial ATA (SATA) disk drives available in various speeds and capacities from Cray. At least one disk drive per chassis is required to house system software; additional disk drives are optional.

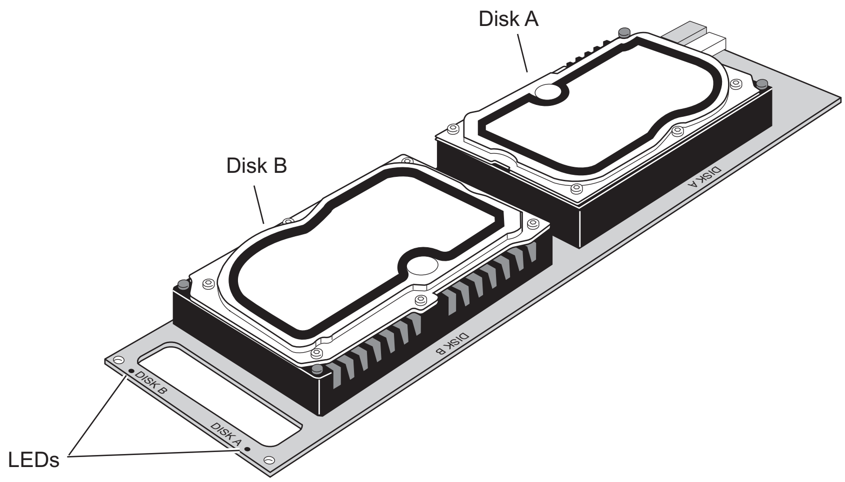

Figure 14. Disk blade with two SATA disk drives

Slots for the disk blades are numbered 1 to 3, from left to right, as viewed from the front of the chassis. The disk blade in slot 1 connects to nodes 1 and 2, the disk blade in slot 2 connects to nodes 3 and 4, and so on. In all cases, disk A (the one farther away from the handle of the disk blade) connects to an odd-numbered node. For instance, for a disk blade installed in slot 3, disk A is for node 5 and disk B for node 6. See Figure 15, page 26 for a logical representation of this mapping.

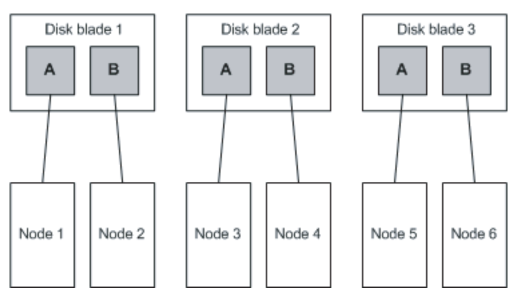

Figure 15. Disk drive connections to nodes

3.13 Power Supply

Each Cray XD1 chassis has one power supply; see Figure 16, page 27.

Single-phase and three-phase models are available. The power switch and status LEDs are on the rear of the chassis; see Section 3.2.7.1, page 14 for a description of the LEDs.

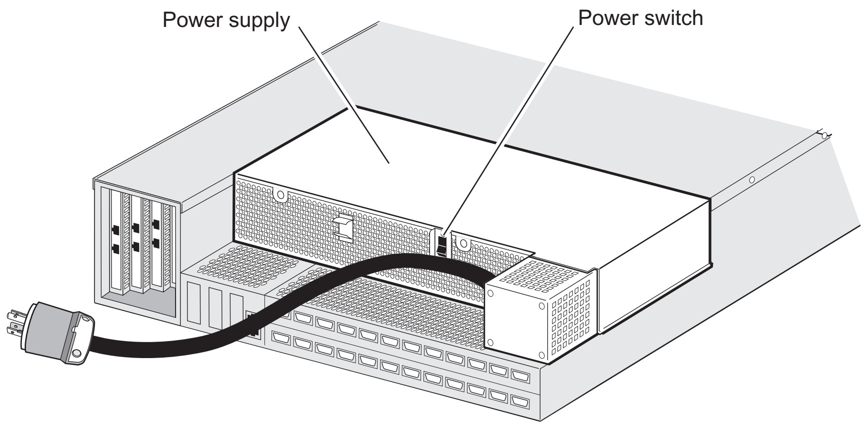

Figure 16. Power supply

Warning: The power switch on the back of the Cray XD1 chassis switches between On and Standby modes. It does not remove power to the chassis. To remove power, disconnect the power cord plug from the power source.

The architecture of the Cray XD1 power supply greatly reduces the impact of failures. The power supply consists of a small collection of shared circuits, and a larger, more complex section of separate power channels, one for each node. The entire power supply goes out of service only if a failure occurs among the collection of shared circuits. If one of the six power channels fails, only the node associated with that channel fails, and the rest of the chassis remains operational. Much of the common Cray XD1 equipment, including the fans, the main board, the management processor, and the RapidArray fabric, draws power from all six channels and so has redundant power.

For Cray XD1 power cord specifications and Cray XD1 power requirements, see the system specifications provided as an appendix to *Cray XD1 Hardware Installation and Upgrade* (HR6-6402).

3.14 Fans

Each Cray XD1 chassis has four field-replaceable fans that are accessible behind the fascias as shown in Figure 17, page 28:

- Three large fans (120 mm by 120 mm) to cool the compute blades, power supply, main board, and fabric expansion card
- One small fan (92 mm by 92 mm) to cool the disk blades and the PCI-X expansion card

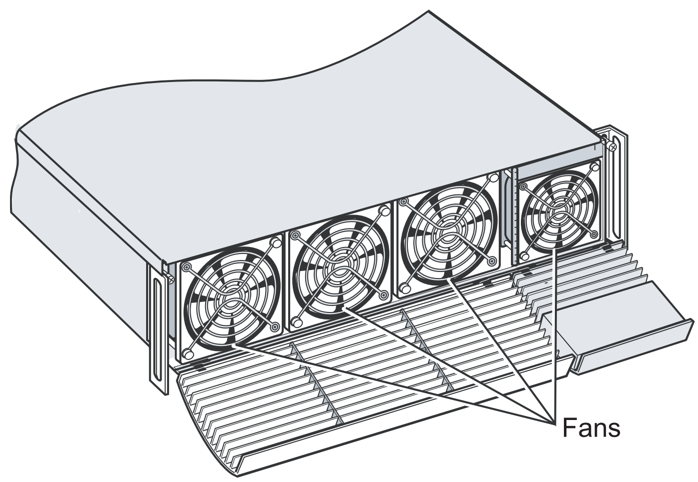

Figure 17. Fans

3.15 Cabinets

Cray provides a four-post cabinet for the Cray XD1 system. It has side walls and is designed to permit enough air to flow through the enclosure to adequately cool system hardware. Full-height (42 VU) and half-height (21 VU) cabinets are available for Cray XD1 chassis. For large systems (24 chassis or more), switch chassis are available for ancillary equipment.

Cray XD1 System Software [4]

This chapter describes the Cray XD1 system software at a high level. For more details, see *Cray XD1 System Administration* (S-2430).

4.1 Linux Operating System

The Linux operating system is broadly accepted in the high performance computing user community for its open source and multivendor interoperability. The Cray XD1 high-performance Linux operating system is based on the SUSE LINUX Enterprise Server (SLES) distribution; see *Cray XD1 Release Description* (S-2453) for the SLES version and the Linux kernel version in the current release. This 64–bit Linux operating system runs on each node.

Note: This document and other Cray XD1 manuals assume that readers are familiar with the Linux operating system. Information on administering and using basic Linux is available from the Linux Documentation Project website (`http://www.tldp.org`) and many commercially available books.

Several features enhance Cray XD1 Linux for HPC purposes. The following sections describe these features in more detail:

- Cray XD1 Linux synchronized scheduler
- MPI support
- Device drivers
- Sockets Direct Protocol

4.1.1 Cray XD1 Linux Synchronized Scheduler

The standard Linux process scheduler was developed to handle interactive user sessions, which are characterized by many independent processes. It is inefficient for large, parallel, compute-intensive applications that require multiple processes to exchange information frequently over a long period of time. Parallel applications generally require that all the processes reach the same point in the program before any of the processes can continue. If even one process is delayed, all processes are delayed.

To address these issues, Cray modified the Linux scheduler. The Cray XD1 Linux synchronized scheduler (LSS) runs all the processes in a parallel job synchronously across all processors, which minimizes the waits for message exchange. For nodes that are dedicated to running parallel jobs, it restricts the time that kernel services and other system daemons consume to a small percentage of the total compute cycle. These enhancements reduce the total execution time for parallel jobs.

4.1.2 MPI Support

A RapidArray Linux device driver, as well as functionality present in the RapidArray processor itself, enable the MPI libraries to interact directly with the RapidArray processor and bypass the kernel. MPI send and receive data goes directly to and from user memory, which reduces memory-to-memory copies and time-consuming kernel context switches.

4.1.3 Device Drivers

Cray XD1 Linux provides all the required device drivers for Cray XD1 hardware, including (but not limited to) the following drivers:

- A driver for the FPGA application acceleration processor provides direct access from a user application to the FPGA and the logic it contains
- A driver for the RapidArray processor enables MPI messages to use the high-speed RapidArray links and supports IP-over-RapidArray communication between nodes
- A driver for the optional Fibre Channel card is automatically loaded when Linux detects that the card is present

4.1.4 Sockets Direct Protocol

Cray XD1 Linux provides an implementation of the Sockets Direct Protocol (SDP) to accelerate applications that rely heavily on interprocessor communication via TCP/IP. Applications that are specified in a system configuration file transparently use SDP over the RapidArray interconnect instead of TCP/IP over RapidArray. Programmers do not need to change source code or even relink applications to take advantage of this feature.

4.2 Lustre File System

The Cray XD1 system runs Lustre, a high-performance, highly scalable, POSIX-compliant shared file system. Lustre uses the Portals lightweight message passing API and an object-oriented architecture to and retrieve data. Lustre file

I/O operations are transparent to the application developer. The I/O functions available to the application developer—Fortran, C, and C++ I/O calls; MPI-I/O calls; and system I/O calls—are converted to Lustre library calls.

Lustre version 1.4, with a custom network abstraction layer for the Cray XD1 RapidArray interconnect, is available as an option for the Cray XD1 system. For more information, see *Lustre File System for the Cray XD1 System* (S-2452).

4.3 Active Manager Software

The Active Manager software monitors and manages a Cray XD1 system. Its architecture, which is described in more detail in *Cray XD1 System Administration* (S-2430), comprises two layers: the infrastructure layer and the application layer.

4.3.1 Infrastructure Layer

The infrastructure layer consists of a single entity that Cray XD1 manuals refer to as the Active Manager *hardware supervisory subsystem*. It runs on the management processor of each chassis and uses the supervisory network to directly monitor and control the hardware. Its responsibilities include maintaining a hardware inventory, starting up and shutting down nodes, monitoring the status of hardware components (including thermal and power characteristics), and taking corrective action when it detects abnormalities. For example, when necessary it controls fan speeds, shuts down nodes, and raises alarms to the Active Manager server.

4.3.2 Application Layer

The application layer is the distributed software that runs on the nodes in each chassis; it is the means by which users interact with the system.

Administrators use the Active Manager application layer to manage and monitor hardware, software versions, job queues, and end-user access, as well as to commission and expand the system. System *partitions* enable administrators to interact with a Cray XD1 system—regardless of its size—as one or more virtual computers.

Administrators divide the pool of available nodes into partitions that support different functions and end-user access rights. For example, some partitions can enable interactive use; others can be reserved exclusively for processing batch jobs. When an administrator creates a partition, the Active Manager software creates a partition master software image for the partition. When the administrator allocates nodes to the partition, the Active Manager software creates a working software image for each one. The node working software images inherit the configuration of the partition master software image. This means that administrators do not have to configure multiple nodes manually. Administrators can also create *custom partitions*, which contain nodes that Active Manager software does not manage.

> **Note:** For a complete list of the Linux configuration files that are managed by the Active Manager software, including details on when the files are updated and which nodes are affected, see the appendixes in *Cray XD1 System Administration* (S-2430).

The Active Manager application layer implemented in Java, and so is portable, transaction-based, and fault tolerant. Its central component is the Active Manager server, a J2EE-based application server that implements the Active Manager business logic. Figure 18, page 33 shows the Active Manager server and other components of the application layer. It shows a typical arrangement for a small system. In this case, all server functions and the data collection function are centralized on a single master node.

Users interact with the Active Manager software through either a command-line interface (CLI) or a browser-based graphical user interface (GUI). As Figure 18, page 33 shows, these user interfaces communicate with the Active Manager server to initiate transactions. The Active Manager server in turn communicates with agents that run on all nodes in the system as well as with the database management subsystem and workload management (WLM) system. Some of the agents communicate with the Active Manager hardware supervisory subsystem (not shown in Figure 18, page 33).

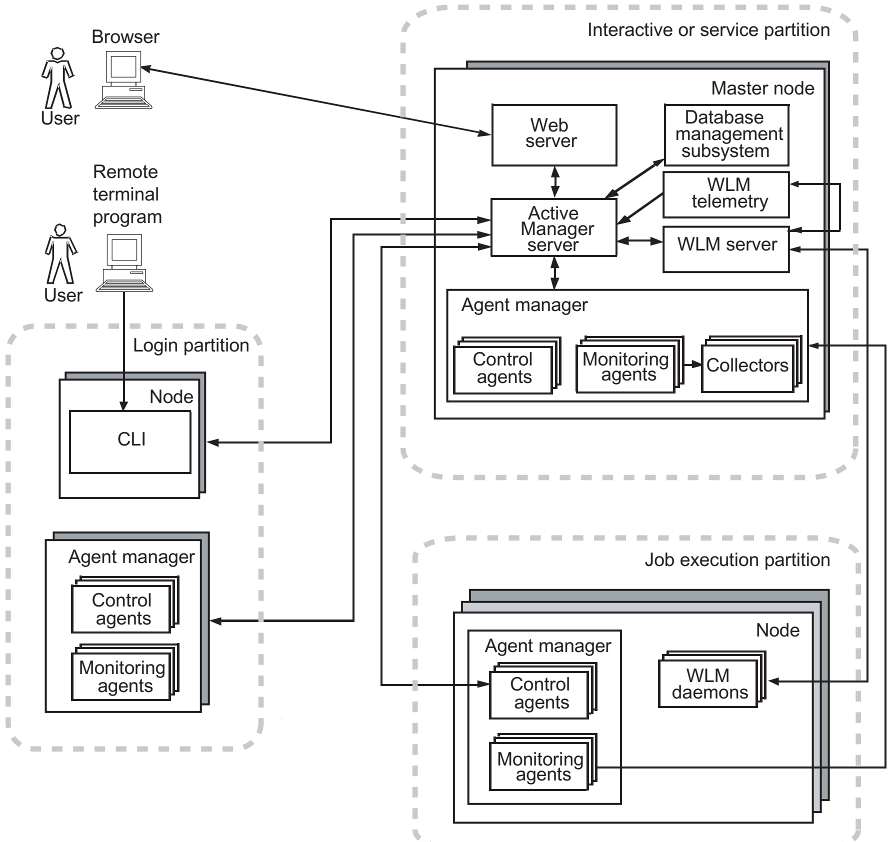

Figure 18. Active Manager architecture: physical view of application layer

4.3.2.1 Active Manager Server

The Active Manager server, a J2EE-based application server, typically runs on a node in a partition that is reserved for interactive use or for services.

The Active Manager server performs the following functions:

- Enables administration and monitoring of the system, partitions, nodes, user access privileges, and jobs

- Manages data stores for system configuration and status
- Coordinates transactions requested by users

4.3.2.2 Web Server

A web server provides a GUI interface to users in the form of web pages. The behavior of the GUI is implemented by Macromedia Flash and by JavaServer Pages, which communicate with the Active Manager server.

4.3.2.3 Agents

While the Active Manager server coordinates all activity in the system, the Active Manager monitoring and controlling agents perform specific tasks at distributed locations. The agents are lightweight remote workers that perform their tasks upon command from the server or that gather data and send it to the server. Most agents are launched by a component called the agent manager that runs on every node—such agents execute as threads of the agent manager process.

4.3.2.4 Database Management Subsystem

The Active Manager software uses a MySQL database, which usually resides on the master node. This is the persistent data store for much of the system configuration and status information that the Active Manager software maintains. In addition, the Active Manager software uses various flat files and log files for certain configuration and status information.

4.3.2.5 Workload Management Systems

The Active Manager software supports two workload management (WLM) systems: Grid Engine and PBS Pro. Either of these WLM systems is integrated with the Active Manager server so that tasks users perform via the Active Manager software configure the WLM system appropriately. In addition, end users can submit, manage, and monitor jobs via the Active Manager GUI.

4.4 Programming Environment

This section describes the Cray XD1 programming environment. For more information on developing code to run on a Cray XD1 system, see the following other Cray publications:

- *Cray XD1 Programming* (S-2433)
- *Cray XD1 FPGA Development* (S-6400)

4.4.1 Scientific and Communication Libraries

The Cray XD1 software distribution includes scientific and communication libraries. For the version numbers of these libraries, see *Cray XD1 Release Description* (S-2453) for the relevant Cray XD1 software release.

4.4.1.1 Scientific Libraries

The Cray XD1 software distribution includes the AMD Core Math Library (ACML) package, optimized for the 64-bit Opteron processor. The ACML package includes the following libraries, available with both FORTRAN 77 and C interfaces. For more details, see the *AMD Core Math Library (ACML)* document, published by AMD.

- Basic Linear Algebra Subprograms (BLAS)—Routines for performing vector-vector, matrix-vector, and matrix-matrix operations; includes Sparse Level 1 BLAS
- Fast Fourier Transform (FFT)—Routines for performing Fast Fourier Transforms and convolutions
- Linear Algebra Package (LAPACK)—Routines for addressing dense linear algebra problems

The Cray XD1 software distribution also includes Scalable Linear Algebra Package (ScaLAPACK)—high-performance linear algebra routines similar to LAPACK, designed specifically for distributed-memory message-passing computers.

4.4.1.2 Communication Libraries

The Cray XD1 software distribution includes the following communication libraries, which consist of routines that can be called from message-passing programs written in C, C++, and Fortran:

- Message Passing Interface (MPI), MPICH implementation—A portable implementation of MPI, a widely-used standard for two-sided communication among processors that run a parallel program
- Shared-Memory Access Library (SHMEM), Generalized Portable SHMEM (GPSHMEM) implementation—Provides a general-purpose shared memory programming interface for parallel programs on a distributed memory system; based on the legacy library introduced by Cray
- Global Arrays (GA)—Provides a shared memory programming interface for parallel programs on a distributed memory system, with support for dense multidimensional arrays and message passing

4.4.2 MPI Programming Model

One of the important features of the Cray XD1 hardware and software architecture is its support for the popular MPI programming model. The following features make a significant difference to the real performance of MPI and other types of parallel job:

- The process synchronization implemented in the Cray XD1 Linux kernel is well suited to jobs that synchronize whenever collective operations are performed.
- The Cray XD1 Linux operating system includes a device driver for the RapidArray processor and so enables messages to make use of the high-speed fabric.
- The MPI library included with the Cray XD1 software is highly optimized for HPC: it performs zero-copy sends and receives and, for short messages, bypasses the Linux kernel.

4.4.3 Application Acceleration Processor Developer's Kit

Cray provides the following implementation tools to help programmers use the FPGA application acceleration processor:

- IP cores—Synthesized logic netlists that also contain placement and timing constraints. They handle the FPGA application acceleration processor's interface to memory as well as to the RapidArray interconnect.
- Reference designs—Sample application logic designs that show completed FPGA implementations.
- Command-line tools—Linux tools that enable users to manually download FPGA logic and access the FPGA application acceleration processor for testing and debugging purposes.
- Application program interface (API)—Programming interface that includes functions to establish communication with the processor, program it, execute the code any number of times for various data, and end communications with the processor. See *Cray XD1 Programming* (S-2433) for more details on the API.
- JTAG interface card—Provides access to the FPGA application acceleration processor's JTAG port.

See *Cray XD1 FPGA Development* (S-6400) for more information.

4.4.4 Development Tools

The Cray XD1 software distribution ships with the GNU toolset, a set of development tools that includes compilers, debuggers, and profiling tools. In addition, the software distribution includes the following performance analysis:

- Performance Application Programming Interface (PAPI)
- CrayPAT
- Apprentice^2

See *Cray XD1 Programming* (S-2433) for more information.

4.5 User Interaction with the System

Administrators and end users use either the Active Manager graphical user interface (GUI) or the command-line interface (CLI) to interact with the system. The browser-based Active Manager GUI provides a user-friendly environment for performing most administrative functions of the system and for submitting and managing jobs. Users can also log in to Linux where they can run standard Linux commands and any additional installed applications to develop programs, edit data, and so on. Administrators and end users who prefer the Linux environment can perform Active Manager functions through the Active Manager CLI.

For more details on the Active Manager user interfaces, see *Cray XD1 System Administration* (S-2430).

Glossary

accepted topology

The fabric components in a present topology that the administrator accepts for logical topology planning.

access privileges

Configuration information that the Active Manager software maintains to control which Linux groups can access the Cray XD1 system. Groups are granted access either to the entire system (administrator privileges) or to one or more partitions.

ACML

AMD Core Math Library

action

In the Active Manager GUI, a simple function that a user performs by clicking a button in the workspace of a page.

Active Manager

The software that monitors and manages all aspects of the Cray XD1 system. Its user interfaces provide administrators and end users with a single point of control for the system.

Active Manager server

A component of the Active Manager architecture; a J2EE-based application server that runs on one or more nodes. The Active Manager server implements the Active Manager business logic.

administrator

A user of the Cray XD1 system with unlimited access privileges, including permission to issue all Active Manager commands. The administrator is responsible for monitoring and managing the system.

agent

A type of component of the Active Manager architecture; a lightweight remote worker that communicates with the Active Manager server to monitor or control one aspect of the system. Some agents run on every node in the Cray XD1 system.

agent manager

An Active Manager process that launches agents as necessary on a node.

AMD Core Math Library

A software package included with Cray XD1 Linux that includes routines for BLAS, FFT, and LAPACK. Routines are available for both Fortran 77 and C interfaces.

application layer

The components of the Active Manager software that run on nodes: the Active Manager server and associated components.

backup node

A node that is physically capable of becoming the master node and is configured to store a periodically synchronized copy of the Active Manager data stores. The administrator can quickly put such a node into service as the master node if the current master node fails or is taken out of service. The enrollment of any node as a backup node is optional; the administrator can enroll one or more suitable nodes, or none.

chassis ID

The permanent numeric identifier of a chassis, unique to each Cray XD1 chassis. A chassis ID has a maximum of six decimal digits.

CLI

Command-line interface.

command-line interface

The set of Active Manager commands that a user can issue from the Linux command line to manage and monitor the Cray XD1 system.

commissioning

The process by which an administrator or installer prepares a Cray XD1 system or component for use after it is physically installed.

compute blade

One of six circuit boards in a Cray XD1 chassis; contains Opteron processors configured as an SMP, DIMMs, and a RapidArray processor. A compute blade may also have an expansion module.

Cray XD1 system

A stand-alone Cray XD1 chassis or multiple chassis that communicate over both the supervisory network and the RapidArray interconnect.

DCP architecture

Direct Connected Processor architecture.

DDR SDRAM

Double data rate synchronous dynamic random-access memory.

direct connect

Also known as "all-to-all mesh." A switchless multichassis topology in which every Cray XD1 chassis in the system connects to every other chassis. A maximum of 13 chassis can be directly connected in a Cray XD1 system. See also *fat tree*.

Direct Connected Processor architecture

A feature of the Cray XD1 system in which the RapidArray interconnect provides high-speed interconnections among all processors in the system, which eliminates the need for slower I/O connection paths.

disk blade

A circuit board of a Cray XD1 chassis that holds one or two SATA disk drives.

disk expansion card

Provides connectors for three disk blades in a Cray XD1 chassis. The disk expansion card is in chassis that do not have the PCI-X expansion card.

end user

A user of the Cray XD1 system who does not have administrator privileges.

expansion module

Optional Cray XD1 hardware that connects to each compute blade; if they are present, a chassis has six expansion modules. The expansion modules provide a node with a second RapidArray processor, two additional Rapid Array links, and an optional application acceleration processor.

expected topology

An XML template that specifies the number of Cray XD1 chassis and switch chassis in a system, whether or not the expansion fabric is present, and the RapidArray links between the chassis. A standard set of templates is provided for the supported system sizes and configurations. An administrator uses the expected topology template to validate a present topology or accepted topology.

fabric

The collection of fabric components that interconnect in the same switching plane. A Cray XD1 system has one or two independently wired, parallel RapidArray fabrics: the main fabric and the optional expansion fabric. These fabrics are also known as fabric X and fabric Y, respectively.

fabric expansion card

Optional hardware in a Cray XD1 chassis that adds a second RapidArray fabric to the system: provides a second internal 24-port RapidArray switch, 12 additional internal links, and 12 additional external ports for chassis interconnection. The fabric expansion card connects to the main board.

fascia

One of two perforated covers on the front of a Cray XD1 chassis. Both are hinged and can be flipped down to provide access to the fans. The right-side fascia also holds the LCD.

fat tree

A physical topology that uses a hierarchy of switches to limit latency and achieve good bisection bandwidth in the system. Cray XD1 systems that exceed 13 chassis are arranged in a fat tree. See also *direct connect*.

Fibre Channel card

Optional hardware in a Cray XD1 chassis that plugs into a PCI-X slot to provide a Fibre Channel interface for a device such as a SAN.

field-programmable gate array

An integrated circuit that consists of arrays of AND and OR gates (typically thousands) that can be programmed to perform complex functions. The Cray XD1 system has optional FPGAs available for use as application acceleration processors.

FPGA

See *field-programmable gate array*.

FPGA application acceleration processor

An FPGA that users can program to accelerate computationally intensive and repetitive algorithms; acts as a co-processor to the Opteron processor. This is an optional component on the expansion module. See also *JTAG interface card*.

hardware supervisory subsystem

The software that runs on the management processor in each Cray XD1 chassis. It primarily monitors the hardware components of the system and proactively manages the health of the system. It communicates with nodes and with the management processors in other chassis over the supervisory network.

infrastructure layer

The Active Manager software components that run on the management processor and programmable components other than nodes: primarily the Active Manager hardware supervisory subsystem. See also *application layer*.

interconnect

See *RapidArray interconnect*.

J2EE

Java 2 Platform, Enterprise Edition.

job

A computing task that runs on one processor or multiple processors concurrently. The workload management (WLM) system assigns the requested resources and launches the job.

JTAG interface card

Optional hardware that tests the application acceleration processors integrated circuits. This card connects to one of the high-speed I/O slots on the main board of a Cray XD1 chassis.

link

See *RapidArray link*.

Linux, Cray XD1

The HPC-optimized Linux operating system, based on the SuSE Linux Enterprise Server (SLES) distribution, that runs on each node. Cray optimizations include the implementation of a synchronized scheduler. See also *synchronized scheduler, Linux*.

logical topology

The set of all fabric paths in a physical topology. It enables any node in a Cray XD1 system to communicate with any other node. See also *accepted topology*, *expected topology*, and *present topology*.

LSS

Linux synchronized scheduler. See *synchronized scheduler, Linux*.

main board

The primary circuit board in a Cray XD1 chassis. The main board has a management processor, RapidArray fabric components, a 10/100 Ethernet switch, three high-speed I/O slots, and connectors for the compute blades, the expansion modules, and the fabric expansion card.

management processor

The processor on the main board of a Cray XD1 chassis that runs the Active Manager hardware supervisory subsystem.

master node

The node on which the Active Manager server runs. See also *backup node*.

Message Passing Interface (MPI)

A widely accepted standard for communication among nodes that run a parallel program on a distributed-memory system. MPI is a library of routines that can be called from Fortran, C, and C++ programs.

node

An instance of the Linux operating system and the hardware components that it controls. The hardware components in a Cray XD1 node include the SMP and its associated memory, one or two RapidArray processors (depending on configuration) and, optionally, an FPGA application acceleration processor. See also *symmetric multiprocessor (SMP).*

node working software image

The software image that is configured for an individual node and from which the node boots; generated automatically by the Active Manager software when the node is allocated to a partition. The node working software image is stored either on the nodes local disk or in the Active Manager repository and NFS-mounted. See also *partition master software image.*

open

For a partition: The setting of the access control attribute of a partition that permits end users to log in or submit jobs, depending on the partitions characteristics. For a node: The setting of the access control attribute of a node that permits end users to log in or submit jobs, depending on the characteristics of its partition.

partition

A logical group of nodes with the same operating system version and configuration; may reside in more than one Cray XD1 chassis. Partitions enable an organization to dedicate a set of nodes to perform a particular function (run a type of job, host a system-wide service, or serve a particular user group). Users treat the set of nodes in a partition as a single, homogeneous computing resource. Administrators specify the attributes of a partition. See also *partition master software image.*

partition master software image

The software image associated with a partition; used to generate the working software images of nodes that are allocated to the partition. The partition master software image is created from a combination of an application release master, a configuration determined by the partitions attributes, any other partition-wide configuration (such as services), and any installed local or third-party software.

PCI-X expansion card

Hardware in the Cray XD1 chassis that provides four PCI-X slots for Gigabit Ethernet and Fibre Channel cards. This card also provides connectors for three disk blades. It connects to one of the three high-speed I/O slots on the main board.

physical topology

The physical arrangement of the RapidArray interconnect in a Cray XD1 system; implemented by cables that connect RapidArray ports on different chassis either directly or via external RapidArray switches. See also *logical topology*.

present topology

All the fabric components that the fabric manager has found through periodic or manually invoked fabric sweeps.

privileges

See *access privileges*.

RapidArray interconnect

The high-speed network that interconnects the nodes in a Cray XD1 chassis, and connects all nodes in a Cray XD1 system via cables and optional external RapidArray switches. The RapidArray interconnect consists of a main and an optional expansion fabric, each with its own set of fabric components. See also *physical topology*.

RapidArray link

The physical communication path between two RapidArray ports. Each link can carry two gigabytes per second.

RapidArray processor

The special-purpose processor on a Cray XD1 compute blade; responsible for most communication functions within the system. The RapidArray processor interfaces an Opteron processor to the RapidArray fabric.

RapidArray switch

A full-crossbar nonblocking switch in the RapidArray fabric. The base configuration includes one 24-port RapidArray switch in each Cray XD1 chassis.

The optional fabric expansion card adds a second RapidArray switch. Equivalent external RapidArray switches are available for implementing fat tree (switched) topologies.

SATA

Serial Advanced Technology Attachment.

service

A unit of work that one software component provides on behalf of another. In the Cray XD1 system, the Active Manager software uses several common system services such as DNS and e-mail transfer.

SMP

Symmetric multiprocessor.

software image

A directory tree that contains the Cray XD1 Linux operating system, application software, and configuration information that is appropriate for the use of the image. See also *partition master software image* and *node working software image*.

supervisory network

The private 100 Mbps Ethernet network internal to the Cray XD1 system that the Active Manager hardware supervisory subsystem uses to monitor and proactively manage the health of the system. Administrators or Cray field engineers can use the supervisory network to commission the Cray XD1 system or diagnose problems.

symmetric multiprocessor (SMP)

In a Cray XD1 system, an SMP is formed from two single- or dual-core Opteron processors and their associated memory. One compute blade holds one SMP. Each chassis contains six compute blades and therefore contains six SMPs. See also *node*.

synchronized scheduler, Linux

The Linux process scheduler customized for the Cray XD1 system. It synchronizes time slots across all nodes in the system and allocates more time slots to computing jobs to maximize application performance. Administrators can configure the allotment of time slots on a partition-by-partition basis.

task

In the Active Manager GUI, a user-initiated function that is complex and possibly requires multiple steps. The user invokes tasks by clicking links in the sidebar of the GUI or in the workspace of the Tasks view, and the Active Manager software guides the user through the steps in the workspace of the page.

view

A section of the Active Manager GUI that relates to one aspect of the Cray XD1 system, such as jobs, partitions, or nodes. Most views consist of multiple web pages.

WLM

Workload management system.

workload management (WLM) system

Software that schedules jobs for execution in a system of networked nodes .

Index